LHC *O*pen *N*etwork *E*nvironment (LHCONE)

Report of the Tier2s connectivity Working Group ("LHCT2S")

Date: March 09, 2011
Version: 2.2

Contents

1 Executive Summary

It is clear from the Bos-Fisk report[1] on Tier 2 (T2) requirements for the Large Hadron Collider (LHC), which this document treats as the definition of the end user requirements, that "unstructured" T2 traffic will be the norm in the future and is already showing up in significant ways on the global Research & Education (R&E) network infrastructures. The R&E network provider community needs to provide ways to manage this traffic, both to ensure good service for the T2s and to ensure fair sharing of the general infrastructure with other uses.

At the same time, it must be recognized that the T2 and Tier 3 (T3) sites will have widely varying needs and capabilities: Some will not use anything but their regular routed IP connections for the foreseeable future and others would happily use a fixed or dynamic lightpath infrastructure if it were available and affordable.

[1] https://twiki.cern.ch/twiki/bin/view/LHCOPN/T2sConn

In response to this evolving environment, CERN convened a group of experts familiar both with the needs of the LHC community and with the design and operation of R&E network infrastructure.

This document is the result of considerable discussion among these experts as to a solution that addresses the functional, technical, and political concerns that must be balanced for a successful solution to the problem.

The result of this effort is a proposal for the “LHC Open Network Environment” (LHCONE). This document is

- A high level architectural description of LHCONE
- Intended to enable different approaches in different continents / regions
- Not intended to be proscriptive; each continent / region will have to figure out how to implement access to LHC ONE
- Not intended to be a complete description of LHC ONE and environment, just LHCONE, at a high level

LHCONE builds on the familiar idea of exchange points – locations and switch fabrics where many networks meet to exchange traffic. MAN LAN, StarLight, and NetherLight are examples of single node policy free R&E exchange points, Atlantic Wave and PacificWave are examples of distributed (multi-node) R&E exchange points, whereas Equinix is an example of a commercial exchange point. LHCONE extends the idea of exchange points to a distributed but integrated collection of inter-connected exchange points that are strategically located to facilitate access by LHC Tier 1 (T1), T2, and T3 sites (collectively referred to as T1/2/3 sites).

The goal of LHCONE is to provide a collection of access locations that are effectively entry points into a network that is private to the LHC T1/2/3 sites. LHCONE is not intended to supplant LHCOPN but rather to complement it. LHCONE is not designed to handle Tier 0 (T0) – T1 traffic. It is anticipated that LHCONE access locations will be provided in countries / regions in a number and location so as to best address the issue of ease of access for the T1/2/3 sites. For example, in the US, LHCONE access locations might be co-located with the existing R&E exchange points and/or national backbone nodes, as much of the US R&E infrastructure is focused on getting R&E institutions to those locations, and those locations are fairly uniformly scattered around the country. A similar situation exists in Europe and Southeast Asia.

As soon as a T1/2/3 site is connected to LHCONE, it can easily exchange data with any other T1/2/3 site over an infrastructure that is sized to accommodate that traffic.

There must be various ways to connect to LHCONE in order to accommodate the varied needs, logistics, and capabilities of the connecting T1/2/3 sites. In particular, LHCONE must accommodate both IP connections and several variations of circuit-based connections. T1/2/3 sites may connect directly or via their network provider (e.g. National Research and Education Network (NREN), US Regional Optical Network (RON), ESnet, etc.).

The design of LHCONE is intended to accommodate the worldwide LHC community and any country or region that wants to set up an exchange point for this purpose can connect that exchange point into LHCONE provided that it meets the service and policy requirements of LHCONE.

In addition to facilitating access for T1/2/3 sites in a way that gets their traffic off of the general R&E IP infrastructure as quickly as possible, LHCONE provides a mechanism to better utilize available transoceanic capacity. As an example, considering the transatlantic R&E paths at present, there is a fair bit of transatlantic capacity that could be available for LHC traffic. However, by using the general R&E infrastructure, which is concentrated on a small number of links, it is difficult to take advantage of this additional capacity. By having LHC T1/2/3 traffic in the purpose-built infrastructure of LHCONE, it is possible to direct the traffic to specific transoceanic paths that, e.g., have spare capacity but that are not used for general R&E traffic. Therefore, in the example of transatlantic inter-connections, LHCONE exchange points in the US and in Europe might use capacity on a significantly more paths than are currently used for T1/2/3 traffic today. Similar situation exists in other regions of the world, such as Asia-Pacific.

In addition to providing data transport, LHCONE provides an infrastructure with appropriate operations and monitoring systems to provide the high reliability (in the sense of low error rates) that is essential for the high bandwidth, high volume data transfers of the LHC community. Further, LHCONE provides a test and monitor infrastructure that can assist in ensuring that the paths from the T1/2/3 sites to LHCONE are also debugged and maintained in the low error rate state needed for LHC traffic. LHCONE does not preclude the continued use of the general R&E network infrastructure by the Tier1/2/3s, as is done today.

This document also contains preliminary thoughts on the policy, governance, funding, and operational stance of LHCONE that enables the services envisioned for the LHC T1/2/3 community.

This document describes LHCONE and how it is going to work for the LHC community. We recognize that the facilities constructed for LHCONE might be of use to other science fields as well.

2 Background

In the 2004/2005 timeframe, the LHC Optical Private Network (LHCOPN) was blueprinted by a small group of experts and managers from the HEP- and (N)REN-communities. This work led to the building, maintenance, and continuous improvement of the LHCOPN as we know it today providing T0-T1 and T1-T1 networking. The LHCOPN is a vital piece of infrastructure in the Worldwide LHC Computing Grid (WLCG). In this model, each T2 and T3 was associated with a T1, in a model that is often referred to as MONARC: The general purpose R&E networks connected the T3s to the T2s and the T2s to T1s in a rather hierarchical and static topology.

In recent meetings and workshops of the LHC community it has become clear that the data models of the experiments are changing to less hierarchical structured ones and that the traffic flows are increasing rapidly. To be more precise, the new data models step away from the static association of T2s to a dedicated T1, and foresee any-T1 to T2, any-T2 to T2, and any-T2 to T3 data transport. At the same time a new model for data placement for access by the analysis programs is emerging: The pre-placing of datasets to specific sites is giving way to a caching model and even a remote I/O model, which may initially reduce network throughput. The breakdown in the MONARC model, i.e. the fact that flows become less predictable, is likely to drive network throughput up in the medium term. Also, as the amount of historical and new data is ever increasing, as the price of storage is decreasing, and as the compute power installed at the sites is ever growing, the data streams themselves are rapidly increasing, possibly posing threats to the service levels of the general purpose R&E networks.

To look into and solve this challenge, CERN took the initiative to ask the community for ideas at the October 2010 LHCOPN Meeting in Geneva. This resulted in four papers with ideas on how to move forward. These ideas were discussed at the January 2011 LHCT2S Meeting at CERN, with people from the HEP- and (N)REN-communities participating. At this meeting it was concluded that the ideas brought forward seem rather compatible with each other and that a core of open exchanges could very well serve as the core of a new infrastructure for T2 networking that would rationalize the Tier 2/3 traffic and permit traffic management.

A small group of experts was tasked to take this outcome to a next level, and prepare a document for discussion on the mailing list before the LHCOPN Meeting in Lyon, France on February 10 and 11, 2011, aiming for an envisaged full consensus with a plan to act.

This document is the result of these efforts and the discussion at the Lyon LHCOPN meeting, and starts with a statement of requirements followed by a proposed, all-inclusive architecture for what is now called the LHC *O*pen *N*etwork *E*nvironment (LHCONE). LHCONE is envisaged to encompass emerging networking technology as it matures, such as multi-domain dynamic lightpaths. This document also discusses stakeholder opportunities and proposes governance and operations models for LHCONE.

3 Data Intensive Science Environment

It is recognized that the LHCONE concept may be of considerable value to other data-intensive science disciplines. Over time, it is possible that LHCONE may become a specific virtual instance on a general data-intensive science open network exchange. If this possibility comes to pass, no policy would exclude use of the physical infrastructure by other science disciplines. However, the LHC community would be provided with a Service Level Agreement that guarantees at least the prescribed bandwidth available to the Tier1/2/3s.

4 LHC End User Environment

The key observations from the Bos-Fisk report regarding the evolution of the end user environment address connectivity, diversity, flexibility, monitoring, and the trend of traffic becoming unstructured.

1. Connectivity

The move away from the strict T0 – T1 – T2 – T3 hierarchy for data management necessitates better connectivity at the T2/T3 level. The T0 – T1 and T1 – T1 network is already well served by the LHCOPN. This not necessarily means a need for higher bandwidth everywhere.

2. Diversity

Three categories of T2/3 sites can be distinguished, among them the top T2s with PBs of storage and many thousands of cores on the floor. Currently the T1s and the top-10 T2s do 75% of all data analysis. For those top T2s we envisage connectivity in the 10 Gbit/s and up range.

On the other side of the spectrum there are sites that currently have too little connectivity to be of general use for the LHC experiments. LHCONE would make their resources available for general LHC experiments' use independent of their size. Inversely it will enable local people to those sites to more ably participate in the analysis of the data. Those sites would be served well with a 1 Gbit/s connection.

All other sites in the middle would need a multiple of 1 Gbit/s links. Although the current contribution to the overall analysis power may be modest, being better connected through LHCONE would greatly enhance their usefulness and their contribution to the overall analysis power, and this is a rather large group of sites, their joint impact is expect to clearly visible.

3. Flexibility

The T0/1 infrastructure does not change very often but it may be expected that T2s and T3s may come and go more frequently. Moreover we now deal with well over 100 sites and political or other changes may influence the overall picture. Therefore LHCONE has to cope with the changes in numbers and locations of the sites.

4. Monitoring

Faultfinding is already difficult on the LHCOPN and will be even more a challenge for this much bigger and more diverse LHCONE network. Monitoring and metering should be part of the design from the start. It needs to be brought into the computing operations rooms of the experiments to allow effective debugging of the whole of the analysis efforts.

5. Trend of Traffic Becoming Unstructured

In the current computing models of the experiments most of the network use is triggered by the central operations people and are reasonably predictable. As the T2/3 traffic will primarily be user analysis driven is must be foreseen that the usage pattern is more chaotic and unpredictable. A couple of thousands users that can access many tens of PetaBytes of data could potentially lead to a lot of network traffic. Moreover this will grow linearly with the amount of available data and the LHC is planned to take data at full capacity for 2011 and 2012.

6. Need to increase T2 – T2 Bandwidth

As the current MONARC model fades in the face of unstructured traffic, one particular short-term chokepoint is the need for more T2 to T2 bandwidth.

7. Need to increase unplanned T2 – T1 Bandwidth.

As the current MONARC model fades in the face of unstructured traffic, a second particular short-term chokepoint is the need for more T2 to T1 intercontinental bandwidth.

5 Design Considerations

Based on the input received, as well as further discussions in the LHCT2S group and at CERN, a high-level, collected list of design considerations was created that sets the boundary conditions for the LHCONE architecture, as follows:

1. LHCONE complements the LHCOPN by addressing a different set of data flows.

   For the time being, LHCONE is physically and operationally distinct from LHCOPN. Over time, LHCONE and LHCOPN may evolve to have common operational components, such as ticketing, when optimizing and looking for synergies.

2. LHCONE enables high-volume data transport between T1s, T2s, and T3s.

   Recent insights from the Bos-Fisk paper indicate that increasingly T2s (and T3s) obtain their data sets from T1s and T2s around the globe, departing from the classical MONARC hierarchical data model. LHCONE enables T2s and T3s to obtain their data from any T1 or T2.

3. LHCONE separates LHC-related large flows from the general purpose routed infrastructures of R&E networks.

   This separation might be accomplished through distinct infrastructure and/or traffic engineering.

4. LHCONE incorporates all viable national, regional and intercontinental ways of interconnecting Tier1s, Tier2s, and Tier 3s.

   The architecture for LHCONE should be inclusive of technology and methods for interconnection that are in general use by R&E networks worldwide.

5. LHCONE uses an open and resilient architecture that works on a global scale.

   T2s and T3s worldwide should be able to join the emerging LHCONE when they are ready, using a method of connecting that fits them best. The core of the LHCONE is built as a resilient infrastructure yielding a very high uptime of the core. We expect T2s and T3s to reach LHCONE via many viable technologies and at different layers.

6. LHCONE provides a secure environment for T1-T2, T2-T2, and T2-T3 data transport.

   The infrastructure provides for private connectivity among the connectors, which might be available along the entire end-to-end path.

7. LHCONE provides connectivity directly to T1s, T2s, and T3s, and to various aggregation networks, such as the European NRENs, GÉANT, and North American RONs, Internet2, ESnet, CANARIE, etc., that may provide the direct connections to the T1s, T2s, and T3s.

8. LHCONE is designed for agility and expandability.

   The architecture of LHCONE should be flexible in dimensions such as accommodating for rising data volumes and the future emergence and adaption of new networking technologies. Also, the architecture of LHCONE should be prepared to accommodate changes in data models of the LHC experiments. In particular, different components may use different network technologies. Also, the components making up LHCONE may evolve over time, and the sites connecting to LHCONE, directly or indirectly, may evolve over time.

9. LHCONE allows for coordinating and optimizing transoceanic data flows, ensuring the optimal use of transoceanic links using multiple providers by the LHC community.

6 Definitions

Tier 1s, Tier 2s and Tier 3s are collectively referred to as "**T1/2/3**."

Any network that provides connections to T1/2/3s, and then in turn connects to LHCONE – e.g. the European NRENs and the North American RONs, and ESnet, etc. – and any network that aggregates aforementioned networks – e.g. the pan-European GÉANT network or the pan-US Internet2 network or the pan-Canadian network CANARIE, is referred to as an "**aggregation network**."

The term "**connector**" refers to any entity that can connect to LHCONE: T1/2/3 and aggregation networks.

The term "**exchange point**" refers to the hardware and physical facilities that provide the access points for LHCONE and the interconnect fabric of LHCONE. From the point of view of the organizations that provide the exchange points, those exchange points themselves may be a distributed exchange point. By "**distributed exchange point**" we understand a geographically distributed collection of network nodes under a single administrative authority, which to a connector appear as one single unit, administratively and operationally.

7 Architecture

The LHC Open Network Environment, LHCONE, builds on the hybrid network infrastructures and open exchange points provided today by the major Research and Education networks on all continents to build a global unified service platform for the LHC community. By its design, LHCONE makes best use of the technologies and best current practices and facilities provided today in national, regional and international R&E networks.

7.1 LHCONE

The LHCONE architecture is based upon the following building blocks:

- Single node exchange points
- Continental / regional distributed exchange points
- Interconnect circuits between exchange points

The continental / regional exchange points are likely to be built as a distributed infrastructure with points of presence (access points) located around the region in ways that facilitate access by the LHC community.

The continental exchange points are likely to be connected by allocated bandwidth on various (possibly shared) links to form LHCONE.

LHCONE is made up of the combination of exchange points and distributed exchange points. These exchange points, and the links in between, collectively provide LHCONE services and operate under a common LHCONE policy.

The architecture allows for an organic growth and modification of LHCONE. Over time, Exchange Points may come (and go) and new Exchange Points can be added to or removed from LHCONE.

7.2 Access methods

The access method is up to the Tier1/2/3s, but may include a dynamic circuit, a dynamic circuit with guaranteed bandwidth, a fixed lightpath, or a connectivity at Layer 3.

We envisage that many of the Tier1/2/3s may connect to LHCONE through aggregation networks.

Figure 1 is intended to illustrate how the building blocks may come together. It is not intended to show how LHCONE is implemented in any given region or imply how any given Tier1/2/3 or aggregation network accesses LHCONE.

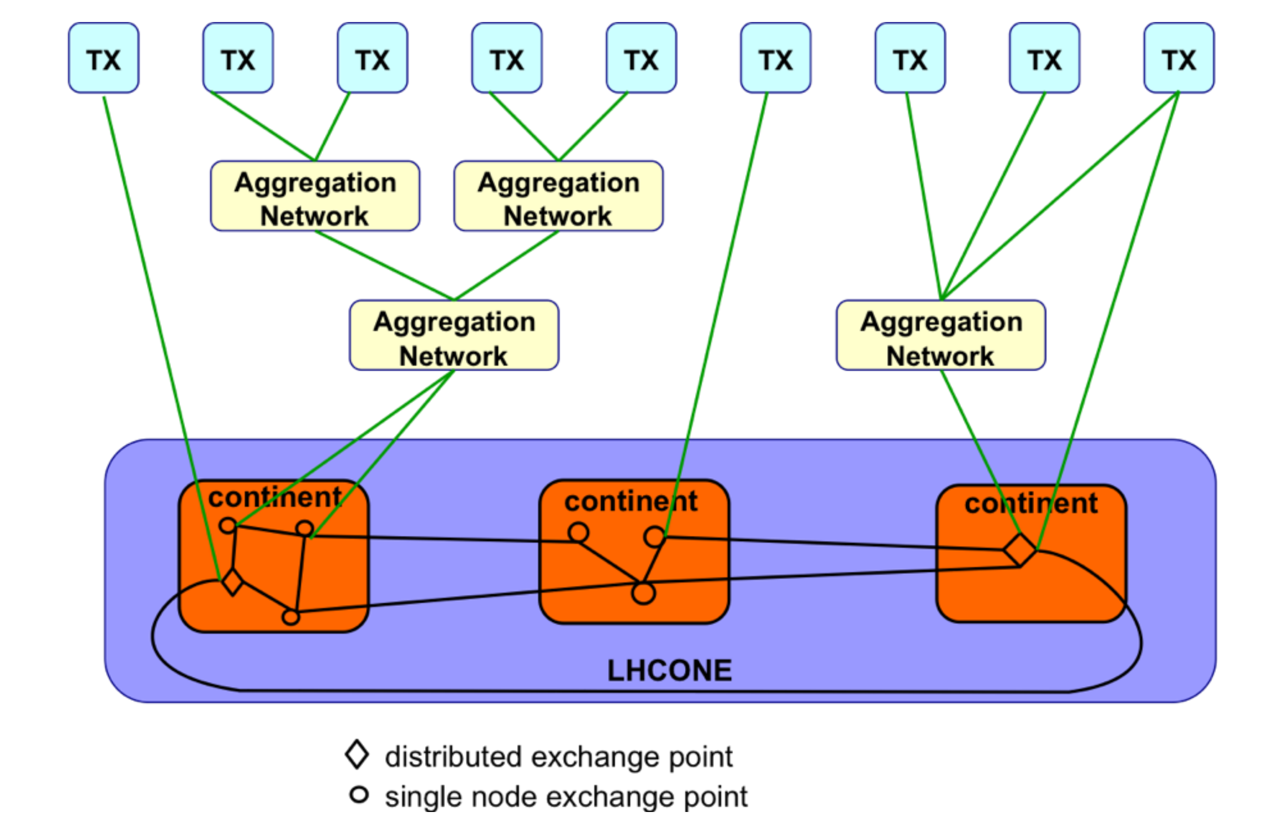

Figure 1. LHCONE Example Implementation for Illustrative Purposes

8 Services

LHCONE is envisioned to provide at least the following basic services to the T1/2/3s:

- Shared Layer 2 domains (private VLAN broadcast domains)
  - LHCONE provides IPv4 and IPv6 addresses on shared layer 2 domains that include all connectors.
  - LHCONE provides private shared layer 2 domains for groups of connectors (Aggregation networks or T1/2/3s) that only want to communicate among themselves.

    - Layer 3 routing is up to the connectors
        - Use of BGP, public IP addresses and public AS numbers is recommended.
        - A Route Server[2] per continent is planned to be available, allowing every site to reach every site if all parties along the path so agree. In order to simplify configuration of the routers of the connector's members can decide to peer only with the Route Servers and get from them all the available prefixes.

- Point-to-point layer 2 connections
    - VLANS without bandwidth guarantees can be set up between pairs of connectors

- Lightpath / dynamic circuits with bandwidth guarantees
    - Lightpaths can be set up between pairs of connectors subject to a resource allocation policy agreed on by the community.
    - LHCONE provides a DICE IDC Version 1.1 protocol compatible circuit management system to facilitate lightpath / dynamic circuit setup[3], with the expectation this will eventually migrate to be compatible with the OGF NSI WG protocol when it emerges.

- A perfSONAR archive provides LHCONE measurements and makes them available, with the expectation this will eventually migrate to be compatible with the OGF NMC WG protocol when it emerges.
    - The presented statistics include current and historical bandwidth utilization values and link availability statistics for any past period of time.
    - LHCONE encourages each Tier1/2/3 and each aggregation network to install a perfSONAR node for both measurement and testing. LHCONE encourages publishing to the LHCONE perfSONAR archive.

The services are available to any LHC computing site, depending on the availability of funding and / or means of connection and the requirements put forward by the site's role in the experiment's computing model.

This list of services is a starting point for LHCONE and not necessarily exclusive.

LHCONE does not preclude the continued use of the general R&E network infrastructure by the Tier1/2/3s, as is done today.

9 Policy

In order for the service to be delivered consistently across LHCONE, it is important to have a consistent policy across the participants. It is expected that the LHCONE policy is defined and may evolve over time in accordance with the governance model defined in a later section.

Our recommended policy is defined below. This should be considered the first draft given to the proposed LHCONE governance.

The policy of LHCONE infrastructure is defined as follows:

[2] A route server is not a router. A route server only distributes route information between the connectors that it peers with, while the routing policy is still being implemented by the connectors. The connectors need to route to any site behind them. A broadcast fabric assumes each connector has a router connected to it in order to be able to reach everyone else on the edge.

[3] Supported at the exchange points by the GLIF Fenius automated GOLE software. This was not discussed at the Lyon meeting, but we expect further discussion on the technical implementation including this point to be presented in v3.0 of this document.

- Any Tier1/2/3 can connect to LHCONE through one or more aggregation networks, and/or exchange points.
- Between the regional exchange points that make up LHCONE, transit is provided to anyone in the Tier1/2/3 community that is part of the LHCONE environment such that they can freely interchange traffic among Tier1/2/3s connected to the LHCONE.
- Exchange points must carry all LHC traffic[4] offered to them (and only LHC traffic), and be built in carrier-neutral facilities so that any connector can connect with their own fiber or using circuits provided by any telecom provider.
- Distributed exchange points must carry all LHC traffic offered to them (and only LHC traffic),and be built in carrier-neutral facilities so that any connector can connect with their own fiber or using circuits provided by any telecom provider and the interconnecting circuits must carry all the traffic offered to them.
- No additional restrictions can be imposed on LHCONE by the LHCONE component contributors.

The scope of this policy framework is restricted to LHCONE. The policies for Tier1/2/3s to connect to aggregation networks are outside the scope of this document. The aggregator networks and/or the Tier1/2/3s might impose additional policy constraints on their own connections. Security on the aggregation networks and the T1/2/3s is the responsibility of the aggregation networks and the Tier1/2/3s and is not the responsibility of LHCONE.

10 Operations

The existing modus operandi in the LHCOPN as well as work on federated operations happening at various locations around the world is the initial guidance for organizing the operations for LHCONE.

11 Implementation Guidance

11.1 Access Switches

Access switches are devices that provide the LHCONE Layer2 Ethernet connectivity with 1G and 10G Ethernet ports; 40G, 100G Ethernet ports are expected to be available in the future. Access switches are part of the exchange infrastructure at those locations were this is available. At other locations, a dedicated switch might be foreseen

11.2 Access links

Access links are Ethernet-framed point-to-point links connecting the connector’s device to one of the LHCONE Access Switches. These links are purchased and operated by the connectors and are not under the responsibility of LHCONE. Any connector may optionally connect to two (or even more) different Access Switches, for resiliency reasons.

12 Governance

Similar to LHCOPN, LHCONE is a community effort; thus a similar governance model is proposed, where all the stakeholders meet regularly to review the operational status,

[4] The scope of LHC sites is defined by LHC experiment MOUs. LHC traffic is defined as traffic between sites fitting within that scope.

propose new services and support models, tackle issues, design, agree and implement improvements.

LHCONE governance defines the policies of LHCONE and requirements for participation. It does not govern the individual participants.

LHCONE governance includes connectors, exchange point operators, CERN, and the experiments, in a form to be determined. It needs to be determined how T2s and T3s that do not connect directly to LHCONE have a voice in governance.

LHCONE governance is responsible for defining how the costs are shared. Costs include, but are not limited to, port costs to connect to LHCONE, the operating and capital costs of LHCONE components, and the operating and capital costs of the links interconnecting the LHCONE components.

LHCONE governance is also responsible for defining how the resources on LHCONE are allocated.

13 Next Steps

In order to achieve the vision of LHCONE, it is necessary to gain operational experience with the proposed approach and an understanding of what might be possible. A formal RFI/RFP-style approach was evaluated and deemed too top heavy. Instead, it is recommended that the community pursue a two-pronged, parallel approach: 1) solicit comments on the proposed approach and 2) implement a bottom-up approach to building a “prototype”[5] that addresses some short-term goals but can likely be built using “existing resources in the R&E community.”

13.1 Short-Term Goals

The goals of this process (in order of importance) include:

1) Evaluating the effectiveness of the proposed LHCONE architecture

Does this architecture support a more versatile approach to data intensive science?

Is this architecture no harder to support operationally?

2) Demonstrating the value of the proposed LHCONE architecture to funding agencies

Does this architecture make more cost efficient use of compute / storage / network resources?

Does this architecture improve the efficiency of scientific discovery?

3) Addressing the short term needs of the LHC Experiments

Augmenting the available T2 – T2 capacity

Augmenting the available transatlantic T2 – T1 capacity

Reducing the load on T1 storage systems

[5] It is anticipated that the “prototype” is more of a cornerstone, in that the experiments are likely to quickly come to depend on it. Therefore, the “prototype” should be expected to remain in place until a replaced by alternative approaches of at least the same utility.

13.2 Architectural Refinement Roadmap

The following roadmap is proposed to bring this architectural document to completion, at least for the near-term. Dependencies are only called out if they are in addition to a dependency on the previous item.

1) Post Architectural Document v2.1 for Public Commentary (Lyon meeting, Calendar Week 6 / Edoardo)

Solicit feedback from Grid Deployment Board (Due Calendar Week 11 / David)

Solicit feedback from funding agencies (Due CW 11)

2) Complete Architectural Document v3.0 (CW 10 / "Small Group")

Expand on architectural definition.

Propose a short-term governance model to cover the prototype implementation period.

Refine prototype use case (dependency on 13.3.1)

3) Send Architectural Document v3.0 to LHCT2S for email approval (CW 10 / Edoardo)

4) Complete Architectural Document v3.1 (CW 13 / "Small Group")

Incorporate public commentary (if any) (dependency on 13.2.1)

5) Send Architectural Document v3.1 to LHCT2S for email approval (CW 13 / Edoardo)

13.3 Prototype Implementation Roadmap

1) Develop Prototype Use Case (CW 9 / Kors Boch & Ian Fisk)

Identify T2 and T1 sites (expected to include 1-2 T2 in Asia, several in North America, and several in Europe, totaling ~10 sites)

Identify BW targets

Identify metrics for success

2) Identify 4 Prototype Team Leads (Now / LHCT2S WG)

4) Identify 4 Prototype Teams (CW 9 / Team Leads)

Assuming the prototype is roughly as identified in 13.3.1, we expect to see:

Team NA: Design an LHCONE component in North America

Team EU: Design an LHCONE component in Europe

Team TA: Identify transatlantic capacity between NA and EU components

Team A: Identify transpacific capacity between Asia and NA component

Team Leads: Coordinate prototype component development

5) Submit 4 prototype components to LHCT2S for approval (CW 13, Edoardo)

6) Implement prototype components and integrate (TBD, Team Leads)

7) Review prototype implementation progress (May 26-27, LHCT2S WG)

8) Evaluate prototype implementation versus short-term goals (TBD, Team Leads / Kors / Ian)

13.4 Beyond the Prototype

Once the prototype is completed and evaluated (likely mid Q3 2011), the LHCT2S WG is expected to develop a follow-on roadmap at a future LHCT2S WG face-to-face meeting. That follow-on roadmap is expected to include:

1) Proposal on more permanent components of LHCONE. The request is likely to call for possible roll-outs of more permanent infrastructure and how it might be funded (some combination of funding agencies and/or costs to connect).

2) Refine the governance model.

3) Refine the service and policy definitions.

4) Refine the architecture.

14 Appendix A: Exchange Point Examples

A classic exchange point like the commercial Equinix exchanges provides an Ethernet fabric and a connection service. Networks typically locate a border router at the Equinix facility and when two networks want to peer they request a VLAN from Equinix. Once connected by the VLAN through the exchange fabric, then the two networks set up a BGP peering between them.

In addition to the sorts of bilateral relationships described above, broadcast VLANs can be set up to allow multiple exchange point connecters to communicate with each other. Layer 2 peering can also be done. The Amsterdam Internet Exchange (AMS-IX), present in five neutral carrier hotels in Amsterdam, is an example of such a large broadcast VLAN in which each all of their 400+ connectors to the AMS-IX has one or more routers present.

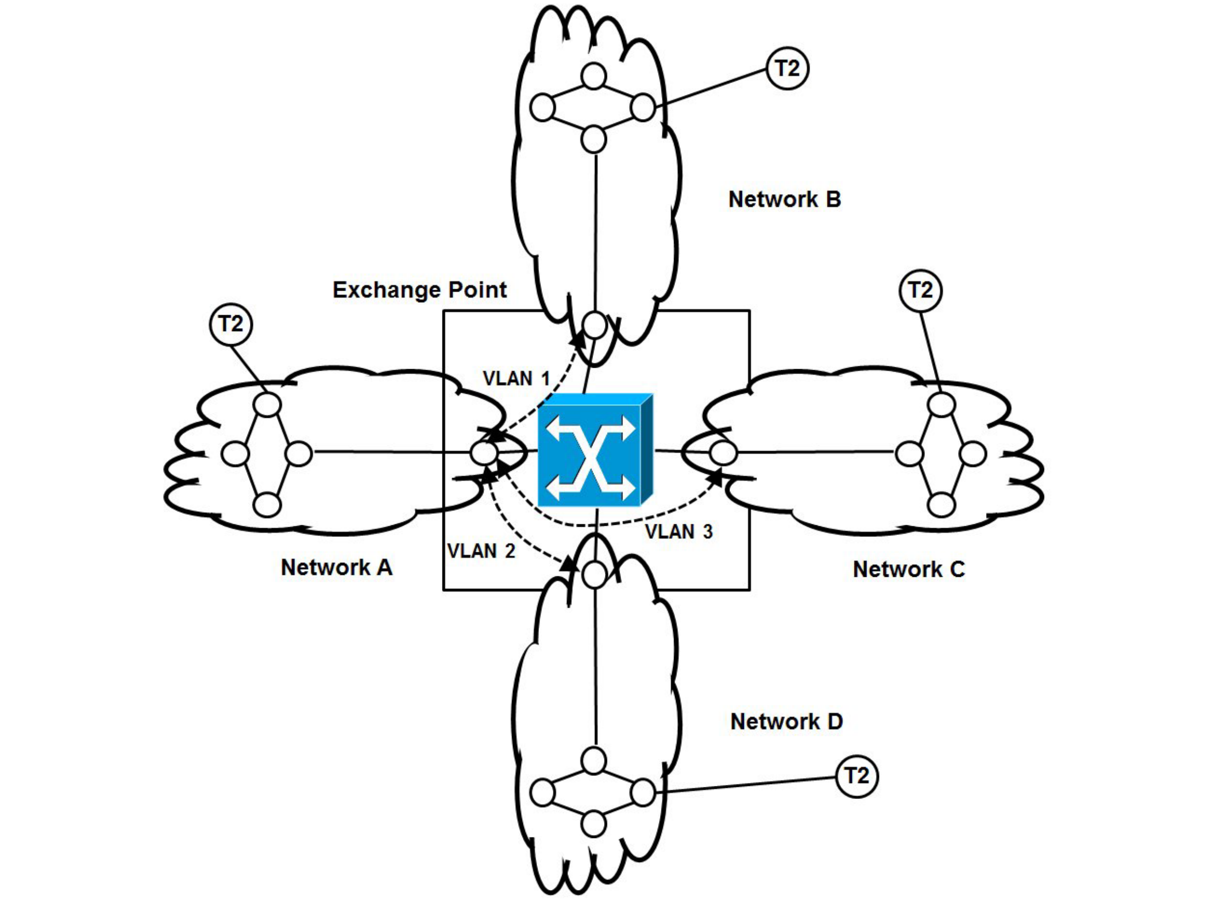

Figure 2. A Layer 2 exchange point

Big exchange points like Equinix have many Ethernet switches in order to accommodate all of the connecting networks, and a “distributed” exchange point would just have some of the switches in different locations. (Figure 3)

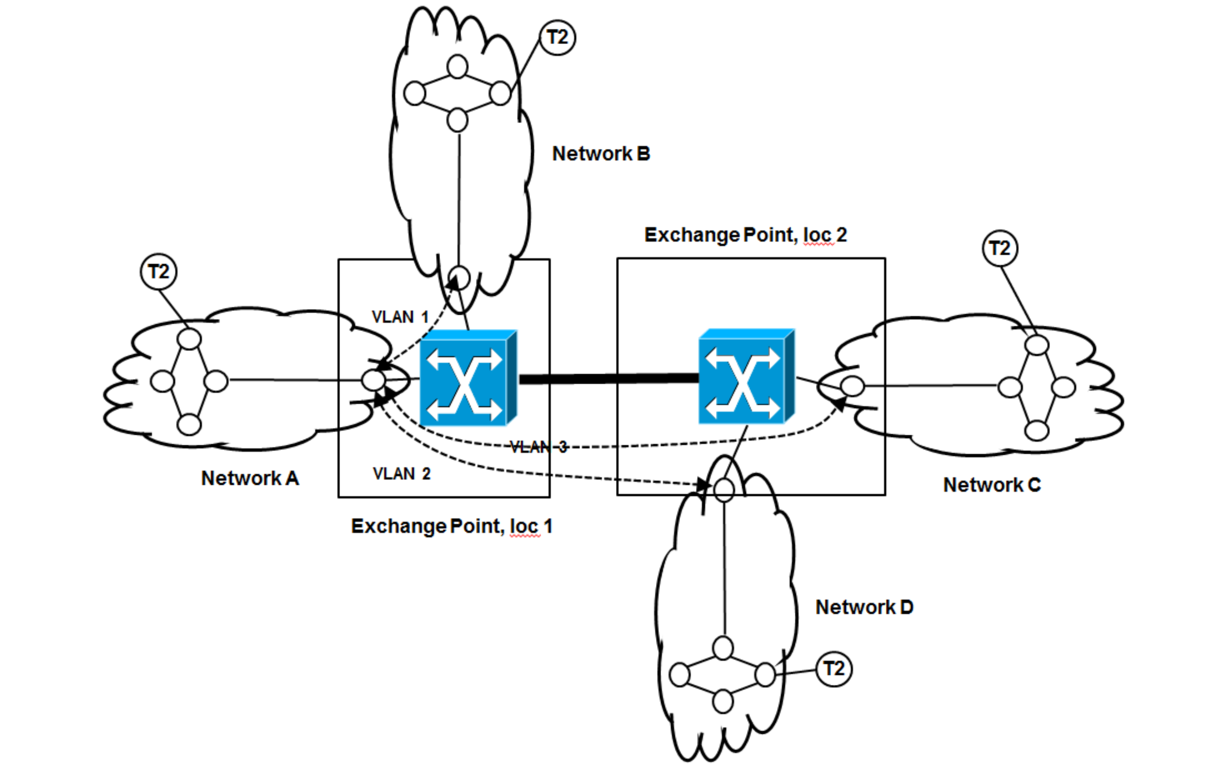

Figure 3. A Layer 2 distributed exchange point

There are several ways that reliability can be provided when the switches of the exchange point are separated by long distances.

One approach to reliability is path redundancy for the interconnecting links. One way to use this is for sites or networks that peer with the exchange point to set up multiple paths over diverse links that are managed by BGP. When one path goes down BGP switches traffic to the other path. This might be accomplished by having the exchange point operator provide two, diversely routed, VLANS (or physical L2 connections) between the networks / sites. This is illustrated in Figure 4.

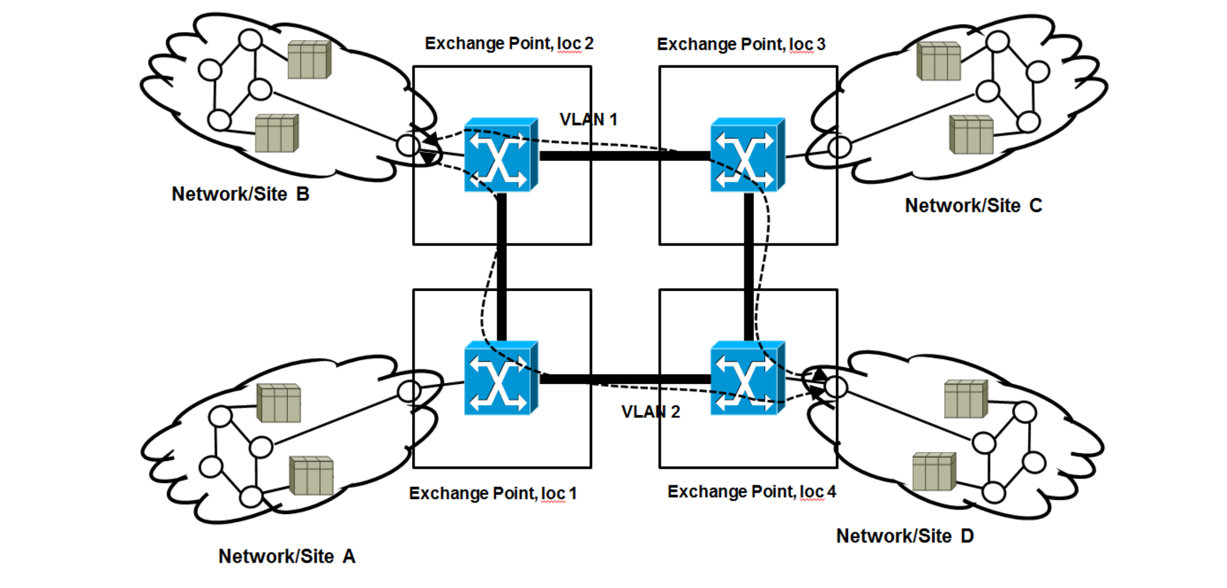

Figure 4. A layer 2 distributed exchange using multiple interconnections for WAN path

diversity

A second approach to exchange point interconnection reliability is to use a lower-level protection such as SONET or optical mesh protection.

The idea is similar to BGP’s management of multiple L2 paths (e.g. failure detection and rerouting); except that the L1 mesh protection happens transparently to the L2 connection between the exchange point switches. Referring to Figure 5, this means that one VLAN that traverses the physical path U-W would be rerouted to physical path U-S-T-W on failure of the U-W path. The layer 2 connection state is maintained apart from a possible change in latency. This approach also provides transparent fail-over of layer 2 connections that are used between connectors as lightpaths.

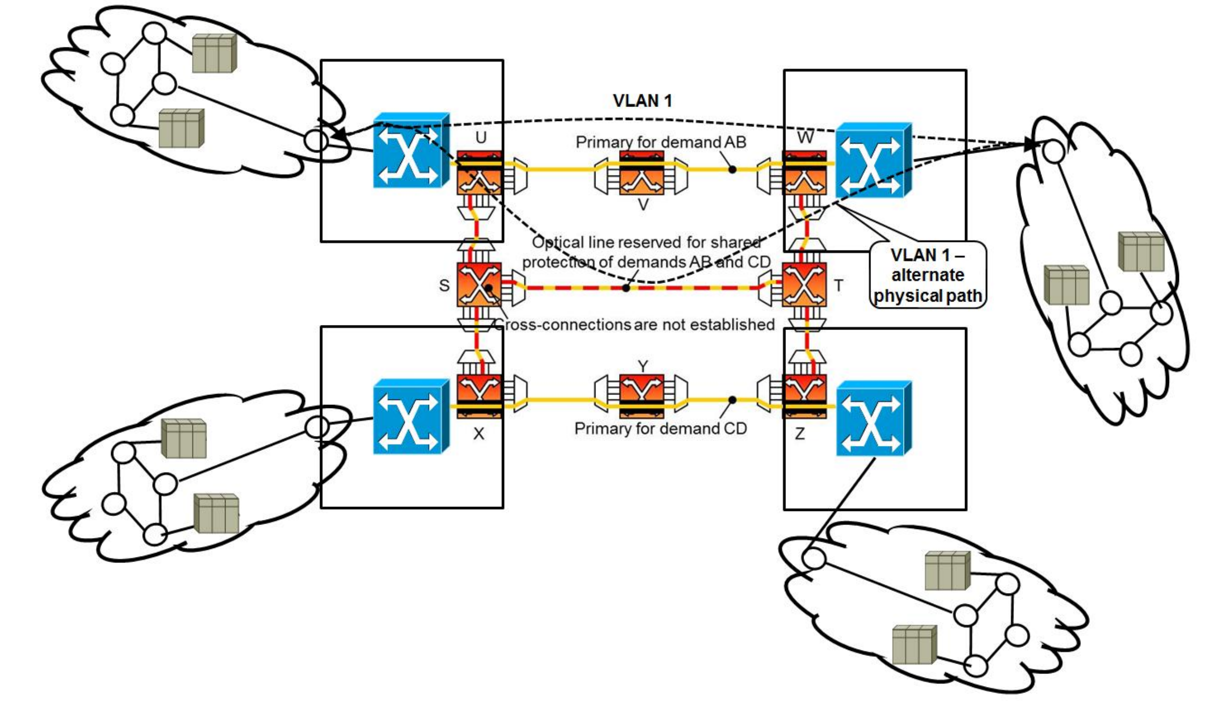
VLAN 1
U
Primary for demand AB
W
V
Optical line reserved for shared protection of demands AB and CD
VLAN 1 – alternate physical path
S
T
Cross-connections are not established
Y
X
Primary for demand CD
Z

Figure 5. A layer 2 distributed exchange using multiple layer 1 paths for reliability (mesh figure from http://en.wikipedia.org/wiki/Optical_mesh_network)

Hybrids of these examples are clearly possible.

15 Appendix B: Previous Work

This document is based on the proposals put forward and discussed during the working group meeting on January 13th in Geneva, as found here: http://indico.cern.ch/conferenceDisplay.py?confId=116636.